JOURNAL OF MULTIVARIATE ANALYSIS **60**, 252–276 (1997)
ARTICLE NO. MV961655

Multivariate Gini Indices

G. A. Koshevoy

C.E.M.I., Moscow, Russia

and

K. Mosler

Universität zu Köln, Cologne, Germany

mean $\mu \in \mathbb{R}^n$; and R_V, related to the expected volume of the simplex formed from $d+1$ independent such vectors. A new characterization of R_D as proportional to a univariate Gini index for a particular linear combination of attributes relates it to the *Lorenz zonoid*. The Lorenz zonoid was suggested as a multivariate generalization of the Lorenz curve. R_V is, up to scaling, the volume of the Lorenz zonoid plus a unit cube of full dimension. When $d=1$, both R_D and R_V equal twice the area between the usual Lorenz curve and the line of zero disparity. When $d>1$, they are different, but inherit properties of the univariate Gini index and are related via the Lorenz zonoid: R_D is proportional to the average of the areas of some two-dimensioned projections of the lift zonoid, while R_V is the average of the volumes of projections of the Lorenz zonoid over all coordinate subspaces.

1. INTRODUCTION

To measure the disparity of a probability distribution, the Gini mean difference and its scale invariant version, the Gini index, are most widely used. The Gini index is closely connected to the Lorenz curve. It amounts to twice the area between the Lorenz curve and the diagonal of the unit square; in other words, it equals the area between the Lorenz curve and its dual. By this, the Gini index is consistent with the Lorenz order: If one Lorenz curve lies below another the Gini index of the first distribution is strictly larger than the index of the second one. Moreover, the Gini index

Received July 18, 1995; revised July 3, 1996.

AMS 1991 subject classification: 62H99, 60E15, 52A21.

Key words and phrases: dilation, disparity measurement, Gini mean difference, lift zonoid, Lorenz order.

0047-259X/97 $25.00

is continuous and scale invariant and, with nonnegative data, has sharp upper bound one.

In this paper we investigate extensions of the Gini mean difference and the Gini index to measure the disparity of a population with respect to several attributes $s = 1, ..., d$. The Gini mean difference of a univariate distribution F is defined as half the expected distance between two independent random variables that both follow the law F. Our first notion is an immediate extension of this. Let $A = [a_{is}]$ be an $n \times d$ data matrix, and a_i its ith row. F_A denotes the d-variate empirical distribution that puts equal mass $1/n$ to each a_i. We define

$$M_{\mathrm{D}}(F_A) = \frac{1}{2n^2 d} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sum_{s=1}^{d} (a_{is} - a_{js})^2 \right)^{1/2} \tag{1}$$

and call M_{D} the distance–Gini mean difference.

The Lorenz zonoid of a univariate distribution is the convex set between the Lorenz curve and its dual. Thus, the Gini index equals the area of this zonoid. The generalized Lorenz curve and its dual form the boundary of another convex set, the lift zonoid, and the Gini mean difference amounts to its area.

The Lorenz zonoid of a d-variate distribution was introduced in Koshevoy and Mosler (1996). It is a convex set in $\mathbb{R}^{d+1}$. The Lorenz zonoid is the scale invariant version of the lift zonoid. For a detailed investigation of these zonoids the reader is urged to consult our two previous articles (Koshevoy and Mosler, 1995, 1996).

Our second notion, M_{V}, is based on the "expanded" volume of the lift zonoid and is named the *volume–Gini mean difference*. It is defined

$$M_{\mathrm{V}}(F_A) = \frac{1}{2^d - 1} \sum_{s=1}^{d} \frac{1}{n^{s+1}} \sum_{1 \leqslant i_1 < \cdots < i_{s+1} \leqslant n} \sum_{1 \leqslant r_1 < \cdots < r_s \leqslant d} |\det(\mathbf{1}, A_{i_1, \ldots, i_{s+1}}^{r_1, \ldots, r_s})|, \tag{2}$$

where $\mathbf{1}$ is a column of ones, and $A_{i_1, \ldots, i_{s+1}}^{r_1, \ldots, r_s}$ is the matrix obtained from the rows $i_1, ..., i_{s+1}$ and the columns $r_1, ..., r_s$ of the data matrix.

For univariate data the Gini index equals the Gini mean difference of the relative data, which are the original data "scaled down" by their mean. For d-variate data we define the distance–Gini index and the volume–Gini index by

$$R_{\mathrm{D}}(F_A) = M_{\mathrm{D}}(\tilde{F}_A) \qquad \text{and} \qquad R_{\mathrm{V}}(F_A) = M_{\mathrm{V}}(\tilde{F}_A), \tag{3}$$

respectively, where F_A is componentwise scaled down to $\tilde{F}_A$ by its mean vector; see Section 3.

Every d-variate Gini index should have at least the following properties: be equal to the usual Gini index in case $d=1$, increase with a proper multivariate extension of the Lorenz order, be scale invariant and continuous, be positive unless the distribution is one-point, and have sharp upper bound one if the data are nonnegative. This and more is shown for our two notions. They are also investigated for general d-variate probability distributions with finite means.

The body of the paper starts with a review of basic features of the univariate Gini index and Gini mean difference, focusing on empirical distributions (Section 2). For univariate distributions, the Gini mean difference increases with the dilation order, and the Gini index increases with the Lorenz order, which we call relative dilation because it amounts to dilation of the relative distributions. Of course, dilation implies relative dilation.

We consider several extensions of dilation to the multivariate case (Section 3). The first is classical d-variate dilation, which means that a random vector X equals another random vector Y plus "noise." The second, directional dilation says that, in each direction p, the projection pX^T equals the projection pY^T plus some "noise" depending on p. X is a directional dilation of Y if and only if the lift zonoid of the distribution of X includes that of the distribution of Y (Koshevoy and Mosler, 1995). Absolute and relative versions of these dilations are considered in Section 3. We show in Section 6 that M_{D} and M_{V} increase with absolute dilation as well as directional absolute dilation. Similarly, both R_{D} and R_{V} increase with relative dilation and directional relative dilation.

Although M_{D} and R_{D} are obvious extensions of the univariate notions, most of their properties have not been explored so far. In Section 4 the distance–Gini index is shown to inherit the main properties of the univariate index plus the *ceteris paribus* property of being proportional to the index of the nondegenerate marginals. In particular, for nonnegative data, R_{D} is bounded by one, and the bound is tight. A surface formula due to Helgason (1980) is used to express $R_{\mathrm{D}}(F_X)$ as proportional to the average over all directions p of $R_{\mathrm{D}}(F_{pX^T})$. It follows from the mean value theorem that, for some specific direction $\tilde{p}$ depending on F_X, $R_{\mathrm{D}}(F_X)$ equals the univariate Gini index of $F_{\tilde{p}X^T}$ times a constant which does not depend on F_X. Using this representation we prove that R_{D} is consistent with relative dilation as well as directional relative dilation. The mean difference M_{D} is investigated in parallel.

Section 5 begins with a theorem relating the volume of the lift zonoid to the expectation of the absolute value of a random determinant. The volume–Gini index is then formed from the volume of the Lorenz zonoid "expanded" by the unit cube. This index is characterized as the average volume of the Lorenz zonoids corresponding to all 2^d-1 marginal

distributions. The volume–Gini index is shown to inherit the properties of the univariate index, to respect the two basic dilation orderings as well as their relative versions, plus the *ceteris paribus* property. Similar properties are derived for the mean difference M_V.

We also establish connections of $M_D(F)$ and $M_V(F)$ with the lift zonoid of F: $M_D(F)$ is proportionate to the average area of certain two-dimensional projections of the lift zonoid (Remark 5.2). $M_V(F)$ is an average volume of projections of the lift zonoid on coordinate planes (Remark 5.3).

Finally, a well known distance-based index is described in terms of L_1 distance, instead of Euclidean distance. All three indices are applied to Fisher's Iris data to corroborate the theory that Iris versicolor is a hybrid of Iris setosa and Iris virginica, in that it shows an intermediate level of diversity from the other two species using each of the indices.

There are several attempts in the literature to define a multivariate Gini mean difference or Gini index. Arnold (1987) proposes $R_D(F_A)$ up to a constant and poses the question of which constant makes it bounded by one. Wilks (1960) introduces the volume of a convex body associated with F. Oja (1983) shows that the Wilks index is the expected volume of a simplex generated by $d+1$ random vertices that are independent and identically distributed according to F; see also Giovagnoli and Wynn (1995). In our framework, the Wilks index amounts to $d+1$ times the volume of the lift zonoid (Theorem 5.1). Torgersen (1991) uses, as a multivariate Gini mean difference, the volume of the zonoid of the distribution, that is, the projection of its lift zonoid on the last d coordinates. For a one-point distribution, both the Wilks–Oja and the Torgersen indices vanish. But also for many other distributions they are zero, which appears to be unsatisfactory. Our notion $M_V(F)$ avoids this drawback; it vanishes if and only if F is a one-point distribution. In addition, we provide the correct scaling factor which makes, for nonnegative data, R_V vary between 0 and 1.

Another multivariate Gini index, associated with a concentration surface, has been introduced by Taguchi (1981). For the relations between Taguchi's concentration surface and the lift zonoid, see Koshevoy and Mosler (1996).

The paper is organized as follows: Some properties of the usual univariate Gini index are surveyed in Section 2. Section 3 presents the definitions of six multivariate dilation orderings and of the lift zonoid and the Lorenz zonoid of a d-variate distribution. Section 4 is about the multivariate distance–Gini index and its properties. The multivariate volume–Gini index is introduced and analyzed in Section 5. In Section 6 we demonstrate that our Gini indices are increasing with multivariate dilations. Section 7 includes the illustrative application to Fisher's Iris data.

Notation. $\mathbb{R}^k(\mathbb{R}^k_+)$ is the k-dimensional Euclidean space of row vectors (nonnegative row vectors). In $\mathbb{R}^k$, x^T is the transpose of a vector x, $\leqslant$ the usual componentwise ordering, and S^{k-1} the unit sphere. $\mathbf{0}$ stands for the origin, and $\overline{x, y}$ for the segment between x and y in $\mathbb{R}^k$. $[a_1, ..., a_l]$ denotes the $l \times k$ matrix with rows $a_1, ..., a_l \in \mathbb{R}^k$. For D and E in $\mathbb{R}^k$, $D \oplus E = \{u: u = x + y, x \in D, y \in E\}$ is the Minkowski sum, and $V_k(D)$ is the k-dimensional volume of D.

2. THE UNIVARIATE GINI INDEX

We shortly survey the Gini mean difference and the Gini index of a univariate distribution. Let $F: \mathbb{R} \to [0, 1]$ be a given probability distribution function on $\mathbb{R}$ that has a finite expectation $\mu(F) = \int_{-\infty}^{\infty} x \, dF(x) \neq 0$.

DEFINITION 2.1 (Gini Mean Difference, Gini Index).

$$M(F) = \tfrac{1}{2} \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| \, dF(x) \, dF(y). \tag{4}$$

is the Gini mean difference of F. $R(F) = M(F)/|\mu(F)|$ is the Gini index of F.

$M(F)$ is the mean Euclidean distance between two independent random variables divided by two, where both random variables are distributed with F. $R(F)$ is the mean Euclidean distance divided by twice the absolute value of the expectation. Let $F^{-1}(s) = \inf\{x: F(x) \geqslant s\}$, $s \in]0, 1]$, denote the inverse distribution function of F, and $GL_F(t) = \int_0^t F^{-1}(s) \, ds$, $t \in [0, 1]$. GL_F is the *generalized Lorenz function*, and its graph is the generalized Lorenz curve of F. The *Lorenz function* is defined by $L_F(t) = \mu(F)^{-1} \, GL_F(t)$, if $\mu(F) > 0$, and $L_F(t) = 1 - \mu(F)^{-1} \, GL_F(1 - t)$, if $\mu(F) < 0$.

The following well known proposition establishes a relation between the Gini index and the Lorenz curve.

PROPOSITION 2.1. *Let* $F(0) = 0$. *Then*

(i) $M(F)$ *equals the area between the graphs of the two functions* $t \mapsto GL_F(t)$ *and its dual* $t \mapsto \overline{GL}_F(t) = \mu(F) - GL_F(1 - t)$, $t \in [0, 1]$.

(ii) $R(F)$ *equals the area between the graphs of the two functions* $t \mapsto L_F(t)$ *and its dual* $t \mapsto \bar{L}_F(t) = 1 - L_F(1 - t)$, $t \in [0, 1]$.

As we demonstrate in Section 5 (Remark 5.1) the assumption $F(0) = 0$ can be dropped, and Proposition 2.1 holds for any F that has finite non-zero expectation.

The special case of an empirical distribution is particularly important. Let F_a denote the distribution function that gives equal weight to each of n, not necessarily different, points a_i in $\mathbb{R}$, $a_1 \leqslant \cdots \leqslant a_n$. Let $a = (a_1, ..., a_n)$ and $\bar{a} = n^{-1}(a_1 + \cdots + a_n)$. Then the Lorenz curve of F_a is the linear interpolation of the points $1/n(k, a_1/\bar{a} + \cdots + a_k/\bar{a})$, $k = 1, ..., n$, in two-space.

$$M(a_1, ..., a_n) = M(F_a) = \frac{1}{2n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} |a_i - a_j| \tag{5}$$

is the *Gini mean difference* of the *sample* $a = (a_1, ..., a_n)$, and

$$R(a_1, ..., a_n) = R(F_a) = \frac{1}{\bar{a}} M(a_1, ..., a_n) \tag{6}$$

is the *Gini index* of a, provided the sample mean is not zero. The Gini index of a equals the Gini mean difference of the "scaled down" sample $\tilde{a} = (a_1/\bar{a}, ..., a_n/\bar{a})$,

$$R(a_1, ..., a_n) = \frac{1}{2n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left| \frac{a_i}{\bar{a}} - \frac{a_j}{\bar{a}} \right|. \tag{7}$$

We state several important properties of the Gini index and the Gini mean difference for empirical distributions of nonnegative data. Parts of them hold as well for general probability distributions and are shown later in the general multivariate case.

PROPOSITION 2.2. (i) *Let* $(a_1, ..., a_n) \in \mathbb{R}^n_+$ *with* $\sum a_i > 0$. *Then*

$$0 = R(\bar{a}, ..., \bar{a}) \leqslant R(a_1, ..., a_n) \leqslant R\left(0, ..., 0, \sum_{i=1}^{n} a_i\right) = 1 - \frac{1}{n} < 1,$$

$$R(\beta a_1, ..., \beta a_n) = R(a_1, ..., a_n) \qquad \text{for every} \quad \beta > 0,$$

$$R(a_1 + \lambda, ..., a_n + \lambda) = \frac{\bar{a}}{\bar{a} + \lambda} R(a_1, ..., a_n) \qquad \text{for every} \quad \lambda > 0. \tag{8}$$

(ii) *R is strictly increasing with the Lorenz order, i.e.,*

$$R(a_1, ..., a_n) > R(b_1, ..., b_n) \qquad \text{if} \quad L_{F_a}(t) \leqslant L_{F_b}(t) \text{ for all } t \text{ and } < \text{ for some } t.$$

(iii) *R is a continuous function* $\mathbb{R}^n_+ \to \mathbb{R}$.

PROPOSITION 2.3. (i) *Let* $(a_1, ..., a_n) \in \mathbb{R}^n_+$ *with* $\sum a_i > 0$. *Then*

$$0 = M(\bar{a}, ..., \bar{a}) \leqslant M(a_1, ..., a_n) \leqslant M\left(0, ..., 0, \sum_{i=1}^{n} a_i\right) = \bar{a}\left(1 - \frac{1}{n}\right) < \bar{a},$$

$$M(\beta a_1, ..., \beta a_n) = \beta M(a_1, ..., a_n) \qquad \textit{for every} \quad \beta > 0,$$

$$M(a_1 + \lambda, ..., a_n + \lambda) = M(a_1, ..., a_n) \qquad \textit{for every} \quad \lambda \in \mathbb{R}.$$

(ii) *M is strictly increasing with the Lorenz order.*

(iii) *M is a continuous function* $\mathbb{R}^n_+ \to \mathbb{R}$.

These and other properties have been investigated by many authors. For surveys and references, see Nygård and Sandström (1981) and Giorgi (1990, 1992).

3. MULTIVARIATE DILATIONS AND THE LIFT ZONOID

Let $\mathscr{F}^d$ be the class of probability distribution functions $F: \mathbb{R}^d \to \mathbb{R}$ with finite mean vector, and $\mathscr{F}^d_0$ be the subclass with finite mean vector no component of which is zero. $\mathscr{F}^d_+ \subset \mathscr{F}^d_0$ denotes the subclass of probability distributions on the nonnegative orthant $\mathbb{R}^d_+$. Given $F \in \mathscr{F}^d$, let $\mu(F) = \int_{\mathbb{R}^d} x \, dF(x) = (\mu_1, ..., \mu_d)$. For every $F \in \mathscr{F}^d$ and $\beta = (\beta_1, ..., \beta_d) \in \mathbb{R}^d$, define $F_{\cdot\beta}(x_1, ..., x_d) = F(x_1\beta_1, ..., x_d\beta_d)$, and $F_{+\beta}(x_1, ..., x_d) = F(x_1 + \beta_1, ..., x_d + \beta_d)$.

For $F \in \mathscr{F}^d_0$, $\tilde{F} = F_{\cdot\mu(F)}$ is called the *relative distribution function*, namely, if F is the distribution function of a random vector $X = (X_1, ..., X_d)$, then $\tilde{F}$ is the distribution of

$$\tilde{X} = \left(\frac{X_1}{\mu_1}, ..., \frac{X_d}{\mu_d}\right).$$

In the sequel, when using $\tilde{F}$, we tacitly assume that $F \in \mathscr{F}^d_0$.

Given F and G in $\mathscr{F}^d$, let X and Y be two random vectors that are distributed according to F and G, respectively. G is a *dilation* of F, $F \preccurlyeq G$, if there exists a random vector Z such that $E(Z \mid X) = 0$ and Y has the same distribution as $X + Z$. The random variable Z may be interpreted as "noise", so that Y is distributed like X plus some noise.

We call G an *absolute dilation* of F, $F \preccurlyeq_a G$, if, $G_{-\mu(G)}$ is a dilation of $F_{-\mu(F)}$. Given F and G in $\mathscr{F}^d_0$, G is a *relative dilation* of F, $F \preccurlyeq_r G$, if, $\tilde{G}$ is a dilation of $\tilde{F}$. For $F \in \mathscr{F}^d$ and $p = (p_1, ..., p_d) \in \mathbb{R}^d$ we denote

$$F(t, p) = \int_{\{x \in \mathbb{R}^d : xp^T \leqslant t\}} dF(x), \qquad t \in \mathbb{R},$$

$$\tilde{F}(t, p) = \int_{\{x \in \mathbb{R}^d : xp^T \leqslant t\}} d\tilde{F}(x), \qquad t \in \mathbb{R}.$$

If F is the distribution function of the random vector X in $\mathbb{R}^d$, then $F(\cdot, p)$ is the distribution function of the random variable $p_1 X_1 + \cdots + p_d X_d$ in $\mathbb{R}$; similarly $\tilde{F}(\cdot, p)$ is the distribution function of $p_1 X_1/\mu_1 + \cdots + p_d X_d/\mu_d$.

G is a *directional dilation* of F, $F \preccurlyeq_{\text{dirr}} G$, if, for every $p \in S^{d-1}$, $G(\cdot, p)$ is a dilation of $F(\cdot, P)$. We say that G is a *directional relative dilation* of F, $F \preccurlyeq_{\text{dirr}} G$, if, for every $p \in S^{d-1}$, $\tilde{G}(\cdot, p)$ is a dilation of $\tilde{F}(\cdot, p)$. Similarly, G is named a *directional absolute dilation* of F, $F \preccurlyeq_{\text{dira}} G$, if, for every $p \in S^{d-1}$, $G(\cdot, p)$ is an absolute dilation of $F(\cdot, p)$.

All these dilations are partial orders (reflexive, transitive and antisymmetric) on $\mathscr{F}^d$, and related by the following implications.

$$\begin{array}{ccc} F \preccurlyeq G & \Rightarrow & F \preccurlyeq_{\text{dir}} G \\ \Downarrow & & \Downarrow \\ F \preccurlyeq_r & \Rightarrow & F \preccurlyeq_{\text{dirr}} G \end{array}$$

$$\begin{array}{ccc} F \preccurlyeq G & \Rightarrow & F \preccurlyeq_{\text{dir}} G \\ \Downarrow & & \Downarrow \\ F \preccurlyeq_a G & \Rightarrow & F \preccurlyeq_{\text{dira}} G \end{array}$$

But, in general, no reverse implication holds. For proofs, see Section 6 below.

Koshevoy and Mosler (1995, 1996) suggested a multivariate generalization of the Lorenz curve and the generalized Lorenz curve that has the following form.

DEFINITION 3.1. Let $F \in \mathscr{F}^d$. For a measurable function $h: \mathbb{R}^d \to [0, 1]$, consider the vector $(z_0(F, h), z(F, h)) \in \mathbb{R}^{d+1}$, where

$$z_0(F, h) = \int_{\mathbb{R}^d} h(x)\, dF(x), \qquad z(F, h) = \int_{\mathbb{R}^d} h(x)\, x\, dF(x).$$

The set

$$\hat{Z}(F) \equiv \{(z_0(F, h), z(F, h)) : h: \mathbb{R}^d \to [0, 1] \text{ measurable}\}$$

is called the lift-zonoid of F. $LZ(F) = \hat{Z}(\tilde{F})$ is called the Lorenz zonoid of F.

The lift zonoid is a multivariate generalization of the generalized Lorenz curve, and the Lorenz zonoid is one of the Lorenz curve. In case $d=1$, the Lorenz zonoid is the area between the Lorenz and the dual Lorenz curves, and the lift zonoid is the area between the generalized Lorenz curve and its dual, (Fig. 1).

Let us recall some properties of the lift zonoid to provide an intuition for working with it. The lift zonoid of $F \in \mathscr{F}^d$ is a convex compact set and contains $\mathbf{0} \in \mathbb{R}^{d+1}$, it is symmetric around $1/2(1, \mu(F))$. If the support of F is in $\mathbb{R}^d_+$, i.e., $F \in \mathscr{F}^d_+$, then $\hat{Z}(F)$ is contained in the $(d+1)$-dimensional rectangle between $\mathbf{0}$ and $(1, \mu(F))$. The lift zonoid uniquely determines the underlying distribution.

The lift zonoid of a probability distribution $F \in \mathscr{F}^d$ can be also seen as the set-valued expectation of the random segment $\overline{\mathbf{0}, (1, X)}$ in $\mathbb{R}^{d+1}$, where

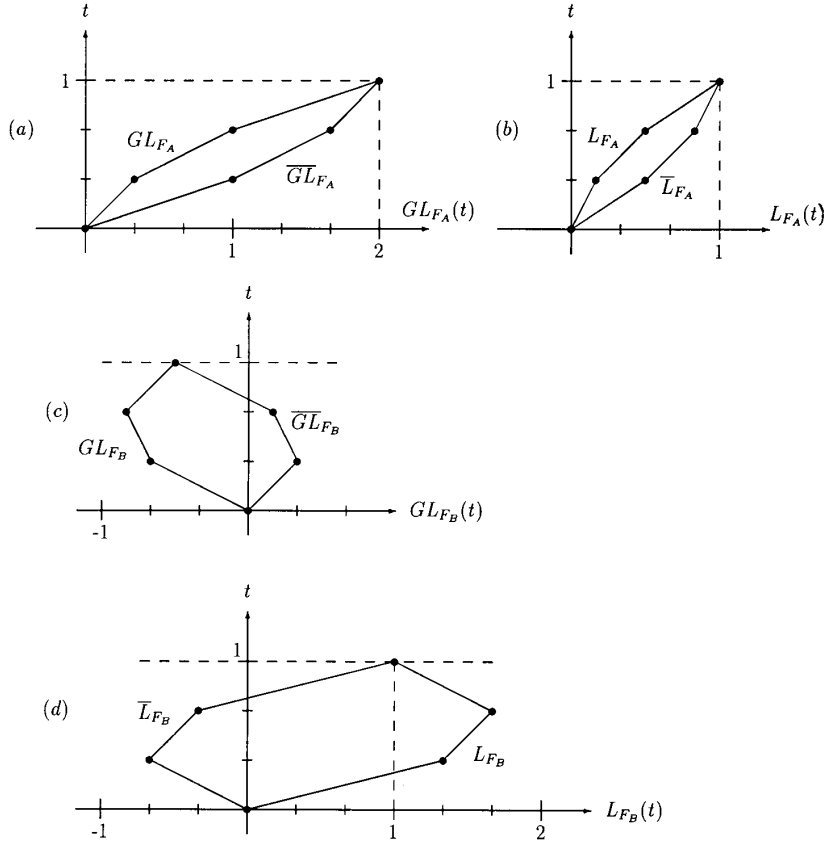

FIG. 1. Lift zonoids and Lorenz zonoids when $d=1$. Lift zonoids (a and c) and Lorenz zonoids (b and d) for F_A and F_B, respectively, where $A=(1, 2, 3)^T$ and $B=(1, -0.5, -2)^T$.

X is a random vector distributed by F. Recall the definition of a random convex set and its expectation. A *random convex set* C is a Borel measurable map from a probability space $(\Omega, \mathscr{B}, P)$ to the space of non-empty, compact, convex subsets of $\mathbb{R}^d$. The *set-valued expectation*, $E(C)$, of a random convex set C is the set given implicitly by

$$\psi_{E(C)}(p) = E(\psi_C(p)), \qquad p \in \mathbb{R}^d, \tag{9}$$

where ψ_C denotes the support function of C. This set-valued expectation has appeared in different settings. See, for example, Weil and Wieacker (1993).

We illustrate the definition of the lift zonoid for empirical distributions. Let $A = [a_{is}]$ be a data matrix, and F_A be the empirical distribution. Then $\hat{Z}(F_A)$ is the Minkowski sum of line segments $\overline{\mathbf{0}, (1/n, a_i/n)}$, $i = 1, ..., n$,

$$\hat{Z}(F_A) = \overline{\mathbf{0}, \left(\frac{1}{n}, \frac{a_1}{n}\right)} \oplus \cdots \oplus \overline{\mathbf{0}, \left(\frac{1}{n}, \frac{a_n}{n}\right)}.$$

$\hat{Z}(F_A)$ is the convex hull of points of the form $\sum_{i=1}^n h_i \cdot (1/n, a_i/n)$, $h_i \in \{0, 1\}$, but not all these points are extreme.

EXAMPLE 1. Let the data matrix be a vector $a \in \mathbb{R}^{\mathrm{d}}$. Then F_A is the empirical distribution putting unit mass to the point a. $\hat{Z}(F_A)$ is the segment that joins $(0, \mathbf{0})$ and $(1, \mathrm{a})$.

EXAMPLE 2. Let the data matrix be

$$A = \begin{pmatrix} 1 & 3 \\ 2 & 2 \\ 3 & 4 \end{pmatrix}.$$

Then F_A is the two-dimensional empirical distribution that puts mass $\frac{1}{3}$ to points $(1, 3)$, $(2, 2)$ and $(3, 4)$. $\hat{Z}(F_A)$ is the convex polytope with the following set of vertices, $\{(0, 0, 0)$, $(\frac{1}{3}, \frac{2}{3}, \frac{2}{3})$, $(\frac{1}{3}, \frac{1}{3}, 1)$, $(\frac{1}{3}, 1, \frac{4}{3})$, $(\frac{2}{3}, 1, \frac{5}{3})$, $(\frac{2}{3}, \frac{4}{3}, \frac{7}{3})$, $(\frac{2}{3}, \frac{5}{3}, 2)$, $(1, 2, 3)\}$.

The set inclusion of lift-zonoids yields an ordering which is equivalent to the directional dilation. The directional relative and absolute dilations are similarly characterized.

THEOREM 3.1. (i) $F \preccurlyeq_{\mathrm{dir}} G$ *if and only if* $\hat{Z}(F) \subset \hat{Z}(G)$,

(ii) $F \preccurlyeq_{\mathrm{dirr}} G$ *if and only if* $LZ(F) \subset LZ(G)$,

(iii) $F \preccurlyeq_{\mathrm{dira}} G$ *if and only if* $\hat{Z}(F_{-\mu(F)}) \subset \hat{Z}(G_{-\mu(G)})$.

For all these properties of the lift zonoid, see Koshevoy and Mosler (1995, 1996).

Both relative dilation and directional relative dilation are multivariate extensions of the usual univariate Lorenz ordering, i.e. the ordering of Lorenz curves. $\preccurlyeq_{\text{dirr}}$ has been named the *multivariate Lorenz order* in Mosler (1994); see also Koshevoy and Mosler (1996). If we compare empirical distributions with the same number, say n, of support points in $\mathbb{R}^d$, dilation and directional dilation correspond to majorization and directional majorization of $n \times d$ matrices; see Marshall and Olkin (1979, ch. 15).

4. THE MULTIVARIATE DISTANCE–GINI INDEX

The definition of the univariate Gini mean difference (4) has the following multivariate generalization.

DEFINITION 4.1. For $F \in \mathscr{F}^d$ the distance-Gini mean difference is

$$M_{\mathrm{D}}(F) = \frac{1}{2d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|x - y\| \, dF(x) \, dF(y) \tag{10}$$

where $\|\cdot\|$ denotes the Euclidean distance in $\mathbb{R}^d$. $R_{\mathrm{D}}(F) = M_{\mathrm{D}}(\tilde{F})$ is the distance-Gini index.

In the case of an empirical distribution function, $M_{\mathrm{D}}(F_A)$ is given by equation (1), and $R_{\mathrm{D}}(F_A)$ by

$$R_{\mathrm{D}}(F_A) = \frac{1}{2dn^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(\sum_{s=1}^{d} \frac{(a_{is} - a_{js})^2}{\bar{a}_s^2} \right)^{1/2}. \tag{11}$$

Several properties of the distance-Gini mean difference and the distance-Gini index follow easily from the definitions. Recall that, for $\beta = (\beta_1, ..., \beta_d) \in \mathbb{R}^d$, we denote $F_{\cdot\beta}(x_1, ..., x_d) = F(x_1\beta_1, ..., x_d\beta_d)$ and $F_{+\beta}(x_1, ..., x_d) = F(x_1 + \beta_1, ..., x_d + \beta_d)$.

PROPOSITION 4.1. *For all* $F \in \mathscr{F}^d$,

(i) $0 \leqslant M_{\mathrm{D}}(F)$,

(ii) $M_{\mathrm{D}}(F) = 0$ *if and only if* F *is a one-point distribution.*

(iii) $M_{\mathrm{D}}(F_{+\beta}) = M_{\mathrm{D}}$ *for all* $\beta_1, ..., \beta_d$.

(iv) M_{D} *is continuous w.r.t weak convergence of distributions.*

PROPOSITION 4.2. *For all* $F \in \mathscr{F}_0^d$,

(i) $0 \leqslant R_D(F)$,

(ii) $R_D(F) = 0$ *if and only if* F *is a one-point distribution.*

(iii) $R_D(F_{\cdot\beta}) = R_D$ *for all* $\beta_1, ..., \beta_d > 0$.

(iv) R_D *is continuous w.r.t weak convergence of distributions.*

Proposition 4.2(iii) says that R_D is *vector scale invariant*, while Proposition 4.1(iii) states that M_D is *translation invariant*. Regarding upper bounds we have the following result.

THEOREM 4.1. *For* $F \in \mathscr{F}_+^d$, *the inequalities*

$$M_D(F) < \frac{1}{d} \sum_{j=1}^{d} \mu_j(F), \qquad R_D(F) < 1,$$

hold and the bounds are sharp.

Proof. Obviously,

$$\begin{aligned} M_D(F) &= \frac{1}{2d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|x - y\| \, dF(x) \, dF(y) \\ &\leqslant \frac{1}{2d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \sum_{j=1}^{d} |x_j - y_j| \, dF(x) \, dF(y) \\ &= \frac{1}{d} \sum_{j=1}^{d} \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x_j - y_j| \, dF(x) \, dF(y) \\ &= \frac{1}{d} \sum_{j=1}^{d} \frac{1}{2} \int_{\mathbb{R}} \int_{\mathbb{R}} |x_j - y_j| \, dF^j(x_j) \, dF^j(y_j) \\ &= \frac{1}{d} \sum_{j=1}^{d} M(F^j). \end{aligned} \tag{12}$$

Here F^j is the jth marginal distribution. As $F \in \mathscr{F}_+^d$ holds, we have $F^j(0) = 0$ for all j and therefore $M(F^j) < \mu_j$, $j = 1, ..., d$. Thus, (12) yields $M_D(F) < 1/d \sum_j \mu_j(F)$.

It is easily seen that the upper bound $d^{-1} \sum_j \mu_j(F)$ cannot be improved. For example, consider the $n \times d$ matrix $A^{(n)} = [a_{is}^{(n)}]$, $a_{is}^{(n)} = n\mu_i(F)$, if $i = s$, $s = 1, ..., d$, and $a_{is}^{(n)} = 0$ otherwise. Then

$$M_D(F_{A^{(n)}}) = \frac{1}{2n^2 d} \left[n \sum_{i=1}^{d} \sum_{j=1,\, j \neq i}^{d} (\mu_i^2 + \mu_j^2)^{1/2} + 2(n-d)\, n \sum_{i=1}^{d} \mu_i \right]. \tag{13}$$

That implies $\lim_{n\to\infty} M_{\mathrm{D}}(F_{A^{(n)}}) = d^{-1}\sum_j \mu_j(F)$, which shows that $d^{-1}\sum_j \mu_j(F)$ is the least upper bound for M_{D}.

The least upper bound for R_{D} is established by passing from F to $\tilde{F}$. Recall that $\mu_j(\tilde{F}) = 1$ for $j = 1, ..., d$. ▌

Consider a property that is desirable for an index of multivariate disparity. It says that, if an attribute is added that does not vary in the population, the disparity index remains essentially unchanged. More precisely, it multiplies by a factor that depends only on the dimension.

DEFINITION 4.2 (Ceteris Paribus Property). Let J^d be a real valued function that is defined on a subset $\mathscr{D}^d$ of $\mathscr{F}^d$, $d \in \mathbb{N}$. We say that J^d, $d \in \mathbb{N}$, has the *ceteris paribus* property if

$$J^{d+1}(F \otimes E_{\xi_0}) = \gamma(d)\, J^d(F) \qquad \text{for all} \quad F \in \mathscr{D}^d,\ \xi_0 \in \mathbb{R},\ d \in \mathbb{N}. \tag{14}$$

Here E_{ξ_0} denotes the univariate one-point distribution at ξ_0, and $\gamma(d)$ is a constant for every d.

THEOREM 4.2. *M_{D} and R_{D} have the* ceteris paribus *property with*

$$\gamma(d) = \frac{d}{d+1}.$$

The proof is obvious from the definition of M_{D}.

THEOREM 4.3. *Let dp denote the rotation invariant area element on the sphere S^{d-1}, $d \geqslant 2$. There holds*

$$M_{\mathrm{D}}(F) = \frac{\Gamma((d+1)/2)}{4d\pi^{(d-1)/2}} \int_{p\in S^{d-1}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |u-v|\, dF(u, p)\, dF(v, p)\, dp, \tag{15}$$

$$R_{\mathrm{D}}(F) = \frac{\Gamma((d+1)/2)}{4d\pi^{(d-1)/2}} \int_{p\in S^{d-1}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |u-v|\, d\tilde{F}(u, p)\, d\tilde{F}(v, p)\, dp. \tag{16}$$

Proof. We use the following formula by Helgason (1980, Lemma 7.2). For every $z \in \mathbb{R}^d$ and $k > 0$ holds

$$\int_{p\in S^{d-1}} |zp^T|^k\, dp = \frac{2\pi^{(d-1)/2}\, \Gamma((k+1)/2)}{\Gamma((d+k)/2)}\, \|z\|^k. \tag{17}$$

From this formula with $k=1$, we conclude that

$$\begin{aligned} M_{\mathrm{D}}(F) &= \frac{1}{2d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-y\|\, dF(x)\, dF(y) \\ &= \frac{1}{2d}\frac{\Gamma((d+1)/2)}{2\pi^{(d-1)/2}}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\int_{p\in S^{d-1}} |xp^T - yp^T|\, dp\right) dF(x)\, dF(y) \\ &= \frac{1}{2d}\frac{\Gamma((d+1)/2)}{2\pi^{(d-1)/2}}\int_{p\in S^{d-1}}\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |xp^T - yp^T|\, dF(x)\, dF(y)\right) dp \\ &= \frac{\Gamma((d+1)/2)}{4d\pi^{(d-1)/2}}\int_{p\in S^{d-1}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, dF(u,p)\, dF(v,p)\, dp. \quad (18) \end{aligned}$$

This proves (15). The result for R_{D} follows immediately with $\tilde{F}$ in place of F. ∎

Recall that the area of S^{d-1} equals $2\pi^{d/2}/\Gamma(d/2)$. Equation (15) in Theorem 4.3 says that the distance-Gini mean difference M_{D} is a constant times the average, over all directions p in the sphere, of the Gini indices of all univariate distribution functions $F(\cdot, p)$,

$$M_{\mathrm{D}}(F) = \frac{\Gamma((d+1)/2)\,\pi^{1/2}}{\Gamma(d/2)\,d}\left[\frac{\Gamma(d/2)}{2\pi^{d/2}}\int_{p\in S^{d-1}} M(F(\cdot,p))\, dp\right], \quad (19)$$

and similarly for $R_{\mathrm{D}}(F)$. Recall, that the Gamma-function $\Gamma(s)=\int_0^\infty t^{s-1}e^{-t}\,dt$ has the following properties: $\sqrt{\pi}=\Gamma(\frac{1}{2})$ and $\Gamma(s+1)=s\Gamma(x)$; and the Beta-function $B(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ is equal to $\Gamma(a)\,\Gamma(b)/\Gamma(a+b)$. Therefore,

$$\frac{\Gamma((d+1)/2)\,\pi^{1/2}}{\Gamma(d/2)\,d} = \frac{\Gamma((d+1)/2)\,\Gamma(1/2)}{2\Gamma((d+2)/2)} = \frac{B((d+1)/2,\,1/2)}{2}.$$

By the mean value theorem we conclude:

COROLLARY 4.1. *For every F there exist some p and $\tilde{p}\in S^{d-1}$ such that*

$$M_{\mathrm{D}}(F) = \frac{B((d+1)/2,\,1/2)}{2}\, M(F(\cdot,p))$$

and

$$R_{\mathrm{D}}(F) = \frac{B((d+1)/2,\,1/2)}{2}\, R(\tilde{F}(\cdot,\tilde{p})).$$

The corollary says that, for every distribution F, there are directions p and $\tilde{p}$ that reflect the dependence structure of F, i.e. the interplay between the attributes, for the Gini mean difference and the Gini index, respectively.

5. THE MULTIVARIATE VOLUME-GINI INDEX

Another view on the univariate Gini index is that it amounts to twice the area between the Lorenz curve and the diagonal. We now extend this view to the multivariate case.

Given $F \in \mathscr{F}^d$, let $X, X_1, ..., X_d$ be independent random vectors each of which is distributed according to F. Q denotes the $(d+1)\times(d+1)$ matrix having rows $(1, X), (1, X_1), ..., (1, X_d)$, and $E\,|\det Q|$ is the expectation of the modulus of its determinant. The term $(d!)^{-1} E\,|\det Q|$ was called a multivariate Gini index by Wilks (1960); see Oja (1983) and Giovagnoli and Wynn (1995). Oja (1983) has interpreted it via the average volume of random simplexes with vertices $X, X_1, ..., X_d$. The following theorem shows that $((d+1)!)^{-1} E\,|\det Q|$ equals the volume of the lift-zonoid of F.

THEOREM 5.1. *Let F be a given distribution function in $\mathbb{R}^d$. Let $X, X_1, ..., X_d$ be independent random vectors each of which is distributed according to F, and let Q denote the $(d+1)\times(d+1)$ matrix having rows $(1, X), (1, X_1), ..., (1, X_d)$. Then*

$$V_{d+1}(\hat{Z}(F)) = \frac{1}{(d+1)!} E\,|\det Q|.$$

Proof. Zonoids are limits of zonotopes. Recall that a zonotope in $\mathbb{R}^k$ is the Minkowski sum of line segments, say

$$\overline{\mathbf{0}, y_1} \oplus \cdots \oplus \overline{\mathbf{0}, y_n} \subset \mathbb{R}^k \qquad \text{with some given} \quad y_i \in \mathbb{R}^k. \tag{20}$$

It has volume (see, e.g., Shephard 1974)

$$\sum_{1 \leqslant i_1 \leqslant \cdots \leqslant i_k \leqslant n} |\det[\, y_{i_1}, ..., y_{i_k}]|. \tag{21}$$

If $k > n$, the volume equals zero, because each determinant has at least two equal columns and therefore vanishes. For a given F, there exists a sequence F_ν, $\nu \in \mathbb{N}$, of distribution functions with finite supports in $\mathbb{R}^d_+$ that converges weakly to F, i.e., $\lim_\nu \int g\, dF_\nu = \int g\, dF$ for every continuous and bounded function $g: \mathbb{R}^d \to \mathbb{R}$. Due to the continuity of zonoids with respect to weak convergence (Bolker 1969), we have $\lim_\nu \delta(\hat{Z}(F_\nu), \hat{Z}(F)) = 0$, where δ is the Hausdorff distance. The volume is a continuous function with respect

to the Hausdorff distance. Therefore, $V_{d+1}(\hat{Z}(F)) = \lim_v V_{d+1}(\hat{Z}(F_v))$. Each volume $V_{d+1}(\hat{Z}(F_v))$ can be calculated by the formula (21). Let F_v have atoms at $x_1, ..., x_m$ with probabilities $q_1, ..., q_m$. Then $\hat{Z}(F_v) = \overline{\mathbf{0}, (q_1, q_1 x_1)} \oplus \cdots \oplus \overline{\mathbf{0}, (q_m, q_m x_m)}$. Hence

$$\begin{aligned} V_{d+1}(\hat{Z}(F_v)) &= \sum_{1 \leqslant i_1 < \cdots i_{d+1} \leqslant m} |\det[(q_{i_1}, q_{i_1} x_{i_1}), ..., (q_{i_{d+1}}, q_{i_{d+1}} x_{i_{d+1}})]| \\ &= \frac{1}{(d+1)!} \sum_{i_1, ..., i_{d+1}=1}^{m} q_{i_1} \cdot \cdots \cdot q_{i_{d+1}} |\det[(1, x_{i_1}), ..., (1, x_{i_{d+1}})]| \\ &= \frac{1}{(d+1)!} E\,|\det Q_{F_v}|. \end{aligned}$$

This completes the proof. ∎

But the volume of a lift-zonoid equals zero rather often, also if F is no one-point distribution. Observe, that if the vectors $x_1, ..., x_n$ are linearly dependent, then the volume of the zonotope in (20) equals zero. Thus, whenever the support of F is contained in a linear subspace of $\mathbb{R}^d$ with dimension less than d, then the volume of the lift zonoid is zero. E.g., in the case of an empirical distribution F, if one of the attributes is equally distributed in the population, or if two attributes have the same distribution, then $V_{d+1}(\hat{Z}(F)) = 0$.

The volume of the Lorenz zonoid is given by the following formula.

$$V_{d+1}(LZ(F)) = \frac{1}{\prod_{j=1}^{d} |\mu_j|} V_{d+1}(\hat{Z}(F)), \qquad F \in \mathscr{F}_0^d. \tag{22}$$

In Mosler (1994) the $(d+1)$-dimensional volume of $LZ(F)$ has been introduced as a multivariate Gini index, called the *Gini zonoid index*. Although this index shows a number of useful properties (boundedness between 0 and 1, 0 at one-point distributions, vector scale invariance; weak monotonicity with multivariate dilations), it may be zero also at distributions that are not concentrated at one point. To avoid this drawback of the Gini zonoid index, we propose the following definition. Let $C^d = \{(z_0, z_1, ..., z_d) \in \mathbb{R}^{d+1} : z_0 = 0, 0 \leqslant z_s \leqslant 1, s = 1, ..., d\}$, which is a d-dimensional cube in $\mathbb{R}^{d+1}$. Instead of the volume of the lift zonoid, we use the volume of the lift zonoid "expanded" by this cube.

DEFINITION 5.1. For $F \in \mathscr{F}^d$, the volume-Gini mean difference is defined by

$$M_{\mathrm{V}}(F) = \frac{1}{2^d - 1} (V_{d+1}(\hat{Z}(F) \oplus C^d) - 1). \tag{23}$$

For $F \in \mathscr{F}_0^d$, $R_{\mathrm{V}}(F) = M_{\mathrm{V}}(\tilde{F})$ is the volume-Gini index.

Equation (23) can be rewritten in the form

$$M_{\mathrm{V}}(F)=\frac{1}{2^d-1}\,(V_{d+1}(\hat{Z}(F)\oplus C^d)-V_{d+1}(\hat{Z}(E_{\mu(F)})\oplus C^d)), \qquad (24)$$

where $E_{\mu(F)}$ is the one-point distribution at $\mu(F)$. (24) says that the volume-Gini mean difference is proportional to the volume of the lift zonoid of a distribution, "expanded" by the d-dimensioned unit cube, minus the volume of the lift zonoid of the one-point distribution at the mean vector, "expanded" by the same cube.

Figure 2 illustrates this in the case $\mathrm{d}=1$.

Remark 5.1. By the lift zonoid approach we are also able to prove a general result on the univariate Gini mean difference. For $d=1$, by Definition 5.1, the volume–Gini mean difference, $M_{\mathrm{V}}(F)$, equals the two-dimensional volume of the lift zonoid, $V_2(\hat{Z}(F))$. Theorem 5.1 says that

$$V_2(\hat{Z}(F))=\frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}}\left|\det\begin{pmatrix}1 & 1\\ x & y\end{pmatrix}\right| dF(x)\,dF(y). \qquad (25)$$

As $|x-y|=|\det(\begin{smallmatrix}1 & 1\\ x & y\end{smallmatrix})|$, we have

$$V_2(\hat{Z}(F))=M_{\mathrm{D}}(F)=M(F).$$

Recall that, for $d=1$, the lift zonoid is the area between the generalized Lorenz curve and its dual. So, we conclude that the usual Gini mean difference equals the area between the generalized Lorenz curve and its dual. This proves that Proposition 2.1 is true not only for distributions on $\mathbb{R}_+$, as stated, but for general ones.

Remark 5.2. $M_{\mathrm{D}}(F)$ and $R_{\mathrm{D}}(F)$ are related to the lift zonoid and the Lorenz zonoid, respectively: For $p=(p_1, ..., p_d)\in S^{d-1}$, let pr_p denote the projection of $\mathbb{R}^{d+1}$ on the two-dimensioned plane that is spanned by the vectors $(1, 0, ..., 0)$ and $(0, p_1, ..., p_d)$. Then, for $z=(z_0, z_1, ..., z_d)\in\mathbb{R}^{d+1}$, we get $pr_p(z)=(z_0, \sum z_i p_i)$ with respect to this base. The projection of the lift zonoid by pr_p equals the lift zonoid of $F(\cdot, p)$ (Koshevoy and Mosler 1995). By this and Corollary 4.1 we can state that $M_{\mathrm{D}}(F)$ is $B(d+1/2, \frac{1}{2})/2$ times the average area of these two dimensioned projections of the lift zonoid. Similarly, $R_{\mathrm{D}}(F)$ is the same with the Lorenz zonoid.

The choice of the constant $1/(2^d-1)$ in (23) is explained in the following theorem. We need some notations: For a nonempty subset $K\subset\{1, ..., d\}$, $F^{(K)}$ denotes the marginal distribution with respect to the coordinates indexed by K.

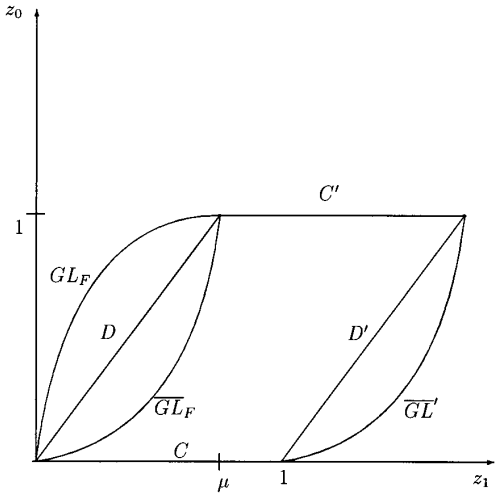

FIG. 2. Definition of M_V when $d=1$. GL_F is the generalized Lorenz curve of a univariate distribution F, and $\overline{GL}_F$ is its dual. The segment D is the lift zonoid of the one-point distribution E_μ at the mean μ of F. GL_F and $\overline{GL}_F$ form the boundary of the lift zonoid $\hat{Z}(F)$. The cube in dimension one is the segment C. Thus, $\hat{Z}(E_\mu) \oplus C$ is the area bounded by D, C, D', C', and $\hat{Z}(F) \oplus C$ is the area bounded by GL_F, $\overline{GL}'$, C, and C'. By (24), the volume-Gini mean difference amounts to the difference of these two areas, which is equal to the area between GL_F and $\overline{GL}_F$.

THEOREM 5.2.

$$M_V(F) = \frac{1}{2^d - 1} \sum_{\varnothing \neq K \subset \{1, \ldots, d\}} V_{|K|+1}(\hat{Z}(F^{(K)})), \tag{26}$$

$$R_V(F) = \frac{1}{2^d - 1} \sum_{\varnothing \neq K \subset \{1, \ldots, d\}} V_{|K|+1}(\hat{Z}(\tilde{F}^{(K)})). \tag{27}$$

Note that Formula (2), M_V for empirical distributions, follows from (21) and (26). For $d=2$, the theorem says that three times $M_V(F)$ equals the volume of the lift zonoid plus the Gini mean differences of the two marginal distributions.

Remark 5.3. By Equation (26), the volume-Gini mean difference is the average of the volumes of projections of the lift zonoid over all coordinate subspaces. They are spanned by $(1, 0, \ldots, 0)$ and $(0, \mathbf{e}_r)$, $r \in K$, $K \in \{1, \ldots, d\}$. Here $\mathbf{e}_r$ is the r-th coordinate unit vector in $\mathbb{R}^d$. By (27) the same holds for the volume-Gini index and the Lorenz zonoid.

Proof of Theorem 5.2. We prove (26) for an empirical distribution F. Then an approximation argument yields (26) for a general distribution. (27) obviously follows from (26).

Let F have atoms at $x_1, ..., x_m$ in $\mathbb{R}^d$ with probabilities $q_1, ..., q_m$. Then

$$\hat{Z}(F) \oplus C^d = \overline{\mathbf{0}, (q_1, q_1 x_1)} \oplus \cdots \oplus \overline{\mathbf{0}, (q_m, q_m x_m)} \oplus \sum_{s=1}^{d} \overline{\mathbf{0}, (0, \mathbf{e}_s)}.$$

Hence, by (21)

$$\begin{aligned} V_{d+1}(\hat{Z}(F) \oplus C^d) &= \sum_{1 \leqslant i_1 < \cdots i_{d+1} \leqslant m} |\det[(q_{i_1}, q_{i_1} x_{i_1}), ..., (q_{i_{d+1}}, q_{i_{d+1}} x_{i_{d+1}})]| \\ &\quad + \sum_{l=1}^{d-1} \sum_{1 \leqslant i_1 < \cdots i_{d+1-l} \leqslant m} \sum_{1 \leqslant s_1 < \cdots s_l \leqslant d} \\ &\qquad |\det[(q_{i_1}, q_{i_1} x_{i_1}), ..., (q_{i_{d+1-l}}, q_{i_{d+1-l}} x_{i_{d+1-l}}), (0, \mathbf{e}_{s_1}), ..., (0, \mathbf{e}_{s_l})]| \\ &\quad + \sum_{i=1}^{m} |\det[(q_i, q_i x_i), (0, \mathbf{e}_1), ..., (0, \mathbf{e}_d)]|. \end{aligned}$$

Let $1 \leqslant l \leqslant d-1$ and $1 \leqslant s_1 < \cdots s_l \leqslant d$ be fixed, $K = \{1, ..., d\} \setminus \{s_1, ..., s_l\}$. Denote x^K the coordinates of a vector x in the set K. Then we have

$$\begin{aligned} V_{|K|+1}(\hat{Z}(F^{(K)})) &= \sum_{1 \leqslant i_1 < \cdots i_{d+1-l} \leqslant m} \det[(q_{i_1}, q_{i_1} x_{i_1}^K), ..., (q_{i_{d+1-l}}, q_{i_{d+1-l}} x_{i_{d+1-l}}^K)]| \\ &= \sum_{1 \leqslant i_1 < \cdots i_{d+1-l} \leqslant m} \det[(q_{i_1}, q_{i_1} x_{i_1}), ..., (q_{i_{d+1-l}}, q_{i_{d+1-l}} x_{i_{d+1-l}}), \\ &\quad (0, \mathbf{e}_{s_1}), ..., (0, \mathbf{e}_{s_l})]|. \end{aligned} \tag{28}$$

In view of $q_1 + \cdots + q_m = 1$,

$$\sum_{i=1}^{m} |\det[(q_i, q_i x_i)(0, \mathbf{e}_1), ..., (0, \mathbf{e}_d)]| = 1. \tag{29}$$

(28) and (29) yield (26). ∎

The following three theorems establish properties of R_{V} and M_{V}.

PROPOSITION 5.1. *For all* $F \in \mathscr{F}_0^d$,

(i) $0 \leqslant R_{\mathrm{V}}(F)$,

(ii) $R_{\mathrm{V}}(F) = 0$ *if and only if* F *is a one-point distribution*,

(iii) $R_{\mathrm{V}}(F_{\beta}) = R_{\mathrm{V}}(F)$ *for all* $\beta_1, ..., \beta_d > 0$.

(iv) R_V *is continuous w.r.t weak convergence of distributions.*

(v) *If* $F \in \mathscr{F}^d_+$, *then* $R_V(F) < 1$ *and the bound is sharp.*

Proof. (i) The volume is a nonnegative function.

(ii) If F is a one-point distribution, then, for every K, $\hat{Z}(\tilde{F}^{(K)})$ is the main diagonal of the unit hypercube in $\mathbb{R}^{\{|K|+1\}}$ and has volume zero. Therefore $R_V(F) = 0$. If F is no one-point distribution, at least one of its univariate marginals, say $F^{(j^*)}$, is the same. Then the univariate Gini index $R(F^{(j^*)})$ is positive. Since $V_2(\hat{Z}(\tilde{F}^{(j^*)})) = R(F^{(j^*)})$, at least one summand in (27) does not vanish, and therefore $R_V(F) > 0$.

(iii) The vector scale invariance is obvious from the definition of $R_V(F)$, since it is based on the relative distribution $\tilde{F}$ only.

(iv) follows from Theorem 7.1 in Koshevoy and Mosler (1995).

(v) For every K, $\hat{Z}(\tilde{F}^{(K)})$ is contained in the unit hypercube of $\mathbb{R}^{|K|+1}$, hence $0 \leqslant V_{|K|+1}(\hat{Z}(\tilde{F}^{(K)})) < 1$, and, by (27), $0 \leqslant R_V(F) < 1$. It is easily seen that the upper bound 1 cannot be improved. For example, consider the distribution $F(x) = \prod_{i=1}^d F_i(x_i)$ where $F_i(x_i) = 0$ if $x_i < 0$, $F_i(x_i) = (n-1)/n$ if $0 \leqslant x_i < 1$, $F_x(x_i) = 1$ if $x_i \geqslant 1$. Then $R_V(F) \to 1$, for $n \to \infty$. ∎

PROPOSITION 5.2. *For all* $F \in \mathscr{F}^d$,

(i) $0 \leqslant M_V(F)$,

(ii) $M_V(F) = 0$ *if and only if* F *is a one-point distribution*,

(iii) $M_V(F_{+\beta}) = M_V(F)$ *for all* $\beta_1, ..., \beta_d$.

(iv) M_V *is continuous w.r.t weak convergence of distributions.*

(v) *If* $F \in \mathscr{F}^d_+$, *then*

$$M_V(F) < \frac{1}{2^d - 1} \sum_{\varnothing \neq K \subset \{1, ..., d\}} \prod_{i \in K} \mu_i \leqslant \frac{1}{2^d - 1} ((\max_i \mu_i + 1)^d - 1)$$

and the first inequality cannot be improved.

The proof is similar to that of Proposition 5.1.

THEOREM 5.3. M_V *and* R_V *have the* ceteris paribus *property with*

$$\gamma(d) = \frac{2^d - 1}{2^{d+1} - 1}.$$

Proof. It is easily seen, that $V_{|K|+1}(\hat{Z}((F \otimes E_\xi)^{(K)}) = 0$ if $d+1 \in K$. If $d+1 \notin K$ then $F^{(K)} = (F \otimes E_\xi)^{(K)}$. This and (26) yield the proposition. ∎

6. CONSISTENCY WITH MULTIVARIATE DILATIONS

The univariate Gini index respects dilation and Lorenz order. We show that our distance-Gini and volume-Gini indices do the same for properly defined extensions of these orderings.

PROPOSITION 6.1 *The following implications hold*

(i) $F \preccurlyeq G \Rightarrow F \preccurlyeq_{\mathrm{r}} G \Rightarrow F \preccurlyeq_{\mathrm{dirr}} G$.

(ii) $F \preccurlyeq G \Rightarrow F \preccurlyeq_{\mathrm{a}} G \Rightarrow F \preccurlyeq_{\mathrm{dira}} G$.

(iii) $F \preccurlyeq G \Rightarrow F \preccurlyeq_{\mathrm{dir}} G \Rightarrow F \preccurlyeq_{\mathrm{dirr}} G$ *and* $F \preccurlyeq_{\mathrm{dira}} G$.

(iv) $F \preccurlyeq_{\mathrm{dirr}} G \Rightarrow R(F(\cdot, p)) \leqslant R(G(\cdot, p))$ *for all* $p \in S^{d-1}$.

Proof. A standard characterization of dilation says that $F \preccurlyeq G$ if and only if $\int \phi(x)\, dF(x) \leqslant \int \phi(x)\, dG(x)$ holds for all convex functions $\mathbb{R}^d \to \mathbb{R}$; see, e.g., the references in Mosler (1994). Further, $F \preccurlyeq G$ implies $\mu(F) = \mu(G)$.

(i) Assume $F \preccurlyeq G$, and let $\phi\colon \mathbb{R}^d \to \mathbb{R}$ be convex. Then, with $(\mu_1, ..., \mu_d) = \mu(F) = \mu(G)$, the function $x \mapsto \phi(x_1/\mu_1, ..., x_d/\mu_d)$ is convex, too. We conclude

$$\int \phi(x)\, d\tilde{F}(x) = \int \phi\left(\frac{x_1}{\mu_1}, ..., \frac{x_d}{\mu_d}\right) dF(x)$$

$$\leqslant \int \phi\left(\frac{x_1}{\mu_1}, ..., \frac{x_d}{\mu_d}\right) dG(x) = \int \phi(x)\, d\tilde{G}(x).$$

Therefore $F \preccurlyeq_{\mathrm{r}} G$. Now assume that $F \preccurlyeq_{\mathrm{r}} G$. Let $p \in S^{d-1}$, $\psi\colon \mathbb{R} \to \mathbb{R}$ convex. Then the function $x \mapsto \psi(xp^T)$ is convex, and from $F \preccurlyeq_{\mathrm{r}} G$ follows that

$$\int \psi(u)\, d\tilde{F}(u, p) = \int \psi(xp^T)\, d\tilde{F}(x)$$

$$\leqslant \int \psi(xp^T)\, d\tilde{G}(x) = \int \psi(u)\, d\tilde{G}(u, p),$$

hence $F \preccurlyeq_{\mathrm{dirr}} G$.

(ii) The proof is similar to that of (i).

(iii) Dilation implies directional dilation. The rest follows from parts (i) and (ii) with $d = 1$.

(iv) If $F \preccurlyeq_{\mathrm{dirr}} G$ and $p \in S^{d-1}$, then $F(\cdot, p)$ is smaller than $G(\cdot, p)$ in relative dilation (= usual Lorenz order). As the usual Gini index is consistent with Lorenz order, we conclude (iv). ∎

Note that, besides the implications given in Proposition 6.1, in general no other implications hold between the various multivariate dilations.

PROPOSITION 6.2. (i) $\preccurlyeq, \preccurlyeq_{\text{dir}}$ *are partial orders* (*reflexive*, *transitive*, *antisymmetric*) *in* $\mathscr{F}^d$.

(ii) $\preccurlyeq_{\text{r}}$ *and* $\preccurlyeq_{\text{dirr}}$ *are preorders* (*reflexive*, *transitive*) *in* $\mathscr{F}_0^d$.

(iii) $\preccurlyeq_{\text{a}}$ *and* $\preccurlyeq_{\text{dira}}$ *are preorders* (*reflexive*, *transitive*) *in* $\mathscr{F}^d$.

Note that the preorders $\preccurlyeq_{\text{r}}$, $\preccurlyeq_{\text{dirr}}$, $\preccurlyeq_{\text{a}}$ and $\preccurlyeq_{\text{dira}}$ are also antisymmetric when applied to the proper factor space.

Proof. (i) The antisymmetry of $\preccurlyeq_{\text{dir}}$ is proven in Koshevoy and Mosler (1995). The antisymmetry of $\preccurlyeq$ follows from the antisymmetry of $\preccurlyeq_{\text{dir}}$ and Proposition 6.1.

(ii) and (iii) follow from (i) and Proposition 6.1. ∎

THEOREM 6.1. *The distance-Gini index* R_{D} *and the volume-Gini index* R_{V} *are strictly increasing with*

(i) *dilation*,

(ii) *directional dilation*,

(iii) *relative dilation*,

(iv) *directional relative dilation.*

Proof. In view of Proposition 6.1, only (iv) has to be shown. Suppose $F \preccurlyeq_{\text{dirr}} G$, hence $R(F(\cdot, p)) \leqslant R(G(\cdot, p))$ for all $p \in S^{d-1}$. Then

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\tilde{F}(u, p)\, d\tilde{F}(v, p) \leqslant \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\tilde{G}(u, p)\, d\tilde{G}(v, p)$$

for all p. Therefore,

$$\int_{p \in S^{d-1}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\tilde{F}(u, p)\, d\tilde{F}(v, p)\, dp$$

$$\leqslant \int_{p \in S^{d-1}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |u-v|\, d\tilde{G}(u, p)\, d\tilde{G}(v, p)\, dp$$

for all p. This yields, according to Proposition 4.3, $R_{\text{D}}(F) \leqslant R_{\text{D}}(G)$. The result for R_{D} follows immediately from Theorems 3.1, 5.2 and the fact that directional relative dilation among two distributions implies the same ordering among their marginals: If $F \preccurlyeq_{\text{dirr}} G$ then $F^{(K)} \preccurlyeq_{\text{dirr}} G^{(K)}$ for all K, $\varnothing \neq K \subset \{1, ..., d\}$. The strict monotonicity is seen from Theorems 3.1,

5.2 and the fact that the lift zonoid uniquely determines the underlying distribution: $\hat{Z}(F) = \hat{Z}(G)$ iff $F = G$ (Koshevoy and Mosler, 1995). ∎

For the distance-Gini and the volume-Gini mean differences, we have an analogous theorem.

THEOREM 6.2. *The distance-Gini mean difference M_D and volume-Gini mean difference M_V are strictly increasing with*

(i) *dilation*,
(ii) *directional dilation*,
(iii) *absolute dilation*,
(iv) *directional absolute dilation.*

Proof. Proofs of (i) and (ii) are similar to those of (i) and (ii) in Theorem 6.1. (iii) and (iv) follow from Propositions 4.1 and 5.2 respectively.

7. CONCLUSIONS

We have investigated two different approaches to extend the usual Gini index and Gini mean difference to the multivariate case. Both extensions preserve important properties of the univariate notions, are increasing with proper multivariate dilations and coincide in the univariate case. They are related via the lift zonoid: the distance-Gini mean difference is proportional to an average of areas of two-dimensional projections of the lift zonoid, the volume-Gini mean difference equals the average of volumes of projections of the lift zonoid on coordinate subspaces. But, in dimensions $d > 1$ they are different. They have the *ceteris paribus* property with different constants. The Gini mean difference is invariant under the transformations of Euclidean space that preserve the distance, i.e., under the orthogonal group. The volume-Gini mean difference is invariant only under the subgroup of reflections. Therefore the two notions can order distributions in opposite directions, as we illustrate by the following example. Consider F_A and F_B with

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 2 + ((7\sqrt{3})/16) & 1 + (7/16) \\ 2 - ((7\sqrt{3})/16) & 1 - (7/16) \end{pmatrix}.$$

Then $\frac{1}{2} = M_D(F_A) > M_D(F_B) = \frac{7}{16}$, while $\frac{1}{3} = M_V(F_A) < M_V(F_B) = 7(1 + \sqrt{3})/48$.

The distance-Gini index and the volume-Gini index of a given empirical distribution are easily calculated. A computer program, written in GAUSS, can be obtained from the authors.

TABLE I

The Multivariate Gini Indices R_D, R_V, and R_S for Three Types of Iris; Data from Fisher (1936). For Further Contrast, the Univariate Gini, Index $R[k]$ Is Given for Each Attribute k, $k = 1, 2, 3, 4$.

	Iris setosa	Iris versicolor	Iris virginica
R_D	0.08536007	0.12217668	0.14415565
R_V	0.042062259	0.067639891	0.083820681
R_S	0.13663000	0.20862000	0.24658000
$R[1]$	0.19620000	0.29000000	0.35136000
$R[2]$	0.20624000	0.17508000	0.17508000
$R[3]$	0.09276000	0.25992000	0.30616000
$R[4]$	0.05132000	0.10948000	0.15372000

Many other multivariate definitions are possible. A popular approach is to use the arithmetic mean, M_S resp. R_S, of the univariate indices,

$$M_S(F_A) = \frac{1}{2n^2 d} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s=1}^{d} |a_{is} - a_{js}|, \tag{30}$$

$$R_S(F_A) = \frac{1}{2n^2 d} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s=1}^{d} \left| \frac{a_{is}}{\bar{a}_s} - \frac{a_{js}}{\bar{a}_s} \right|. \tag{31}$$

This is tantamount to employing the L_1 distance instead of the Euclidean distance in our distance–Gini notions. It can be shown that always $R_D(F) \leqslant R_S(F)$ and $R_V(F) \leqslant R_S(F)$ hold. But this approach, as the index depends on the marginals only, does not reflect the dependency structure of the underlying distribution.

To illustrate our notions, we calculate them for R. A. Fisher's Iris data (Fisher 1936). The data include the measurements of four attributes, sepal length and width and petal length and width, of fifty plants for each of three types of Iris, *Iris setosa*, *Iris versicolor* and *Iris virginica*. The data have been used to test the hypothesis that Iris versicolor is a hybrid of the two other species.

As we can see from Table I, the four attributes are most variable at different types of Iris, as measured by their univariate Gini indices. E.g., the first attribute, petal length, varies most with Iris virginica, while the second attribute, petal width, has its maximum Gini index with Iris setosa. But the three multivariate Gini indices, R_D, R_V, and R_S, order the variability of the three samples in the same way,

$$\text{Iris setosa} < \text{Iris versicolor} < \text{Iris virginica}.$$

Under the assumption that a hybrid has intermediate variability, we conclude that all three multivariate Gini indices back the hypothesis that Iris versicolor is a hybrid of the two other species.

ACKNOWLEDGMENTS

We are grateful to two referees for their hints to improve the presentation. To one of them we owe parts of the introduction. We thank Stephan Erkel for his comments on a previous version and Ulrich Casser for writing the computer program and calculating the numerical example.

REFERENCES

1. Arnold, B. C. (1987). *Majorization and the Lorenz Order*: *A Brief Introduction*. Springer-Verlag, Berlin.
2. Bolker, E. D. (1969). A class of convex bodies. *Trans. Amer. Math. Soc.* **145** 323–346.
3. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, No. 2, 179–188.
4. Giorgi, G. M. (1990). Bibliographic portrait of the Gini concentration ratio. *Metron* **48** 183–221.
5. Giorgi, G. M. (1992). *Il rapporto di concentrazione di Gini*. Libreria Editrice Ticci, Siena.
6. Giovagnoli, A., and Wynn, H. P. (1995). Multivariate dispersion orderings. *Statist. Probab. Lett.* **22** 325–332.
7. Helgason, S. (1980). *The Radon Transform*. Progress in Mathematics, Vol. 5. Birkhäuser, Boston.
8. Koshevoy, G. A., and Mosler, K. (1995). A geometrical approach to compare the variability of random vectors. In *Discussion Papers in Statistics and Quantitative Economics*, Vol. 66. UniBw Hamburg.
9. Koshevoy, G. A., and Mosler, K. (1996). The Lorenz zonoid of a multivariate distribution. *J. Amer. Statist. Assoc.* **91** 873–882.
10. Marshall, A. W., and Olkin, I. (1979). *Inequalities*: *Theory of Majorization and Its Applications*. Academic Press, New York.
11. Mosler, K. (1994). Majorization in economic disparity measures. *Linear Algebra Its Appl.* **199** 91–114.
12. Nygård, F., and Sandström, A. (1981). *Measuring Income Inequality*. Almqvist and Wiksell, Stockholm.
13. Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1** 327–332.
14. Shephard, G. C. (1974). Combinatorial properties of associated zonotopes. *Canad. J. Math.* **26** 302–321.
15. Taguchi, T. (1981). On a multiple Gini's coefficient and some concentrative regressions. *Metron*, 69–98.
16. Torgersen, E. (1991). *Comparison of Statistical Experiments*. Cambridge University Press, Cambridge, MA.
17. Weil, W., and Wieacker, J. A. (1993). Stochastic Geometry. In *Handbook of Convex Geometry* (P. M. Gruber and J. M. Wills, Eds.), North-Holland, Amsterdam, 1391–1438.
18. Wilks, S. S. (1960). Multidimensional statistical scatter. In *Contributions to Probability and Statistics in Honor of Harold Hotelling* (I. Olkin *et al.*, Eds.), Stanford Univ. Press, Stanford, CA, 486–503.